Group Project

STAT 300 (Spring 2025)

A used car business owner (Rohan) in an Asian country is trying to develop a model to understand the features of used cars that impact the sale price. He has collected data on 90 cars, and it is given in *usedcars.csv* file. Below are the variable description.

carID - A unique identification number of the car

KmDriven - The total kilometers driven by the car

FuelType - The type of fuel used by the car (Petrol or Diesel)

Transmission - The transmission type of the car (Manual or Automatic)

Mileage - The fuel efficiency of the car in kilometers per liter

Engine - The engine capacity of the car in Cubic Centimeters.

Power - The maximum power output of the car in bhp

Seats - The number of seats available in the car

Age - The age of the car (this year - manufactured year)

Price - The selling price of the car in rupees

Rohan is seeking your assistance to *build a model to make inferences and predictions about sale prices.*

That is, the research questions to be answered in your project are:

1. Are any of the features measured useful in predicting the sale prices of used cars? If so, what are they and how do they affect the the sale price of used cars?
2. How can the measured features be used to predict the sale price of the used cars?

As a group, you are expected to perform a statistical analyses to answer Rohan's questions and prepare a summary report.

(1). First you are going to split the dataset into two: training and testing set. Use 72 observations in the dataset to train your models and remaining 18 to validate the prediction performance. When subsampling 72 observations for the training data, I want you to use a random seed. This is for the purpose of reproducibility of the results. This random seed MUST have the format "300GroupNo". For instance, if your group is 1, the your random seed should be 30001. If your group number is 11 then, your random seed should be 30011, etc. Remember, for descriptive analysis as well as for developing the model, you are going to use your training set only. You will use the testing set to validate the prediction performance of the developed model.

You may adapt a code like below to split the data into testing and training set:

```
#Read the data
ucars<-read.csv("usedcars.csv")

#Set the seed following the format I have given above.
#For instance if you are in group 1 your seed should be 30001,
#if you are in group 12 your seed should be 30012.
#But I set it to 1234 as an example

set.seed(1234)

#Decide which records to keep in the training set. I sort them for the convenience
trainRec<-sort(sample(1:nrow(ucars),size=72))
#Extract the indexes that were not in trainRec to the test records
testRec<-setdiff(1:nrow(ucars),trainRec)

#Extract indexed rows in trainRec and keep them in the training set
traincars<-ucars[trainRec,]
#Extract indexed rows in testRec and keep them in the training set
testcars<-ucars[testRec,]
```

(2). Describe the data using training set. This include any behaviors of the individual variables, potential relationships that you can observe from graphical and numerical summaries of the data.

(3). Use appropriate variable selection strategies to develop a model to illustrate the features that affect the sale price of used cars. You must verify the validity of the model assumptions and utilize appropriate transformations to remedy any deviations you might observe. To pick the model, you are going to use your training set. You must clearly justify the reasons for your suggested model

(4). Based on all your observations, you have to recommend one model to your client. Clearly identify the important variables when it comes to determining the sales based on your developed model.

(5). Finally, you must also illustrate the prediction ability of the developed model using testing set. Think of creative ways to convey this to your client (Rohan).

As a group, you are expected to perform a statistical analyses to achieve the aforementioned objectives and prepare a summary report. Your report should contain your data analysis steps to cover above; how you arrived at your final model should be clearly explained. In doing so, you need to select which graphs and tables are essential to be included in the report. All figures and tables must be labeled and hyperlinks should be utilized when referring to the relevant tables/graphs. When you include a table, it must be a formatted table, not an R output. You should include the corresponding output in the appendix of the report.Be creative in your final remarks.

The main text of your report should be formatted to a **maximum of 4 pages** (standard letter size), 1.5 line spacing with 1 inch margin. Font type must be 10pt-Arial.
Select most important figures and tables to be presented. Any additional methods, tables and figures can be submitted as an appendix. Appendix does not have a page limit. Make sure to acknowledge any published resources you used, and your reference list can be outside the 4-page limit as well. Your reports should be submitted as a single PDF file. **Please note that only the first 4 pages of your report will be graded and appendix will be only be referred in special circumstances. All main results should be presented within the 4-page limit.**

Your short report should include:

- within the 4-page limit:
    - Background and Introduction (what is the research question you are trying to answer and why is it interesting, relevant information about the data, reproducibility

of the results, etc.)

- Exploratory data analysis/descriptive analysis: Include any gradpical and numerical summaries. Just including graphs/tables are not sufficient. You must explain what insight they give you into the statistical analysis.
- Statistical Methods and Analysis (how you arrived your final model from the variable selection approaches,any transformations, your observations regarding assumptions, validation of the predictions using test set, etc.)
- Conclusion (Your final conclusions/recpmmendations summarized based on the statistical analysis)
- Discussion (limitations, future directions, etc.)

- outside the page limit:
  - References
  - Appendix (if needed): any additional figures and tables, methods, etc.

Everyone in the group **MUST** put equal amount of time and effort to the project and it is your responsibility to let me know if someone in your group is not participating. Each group **MUST** submit a document signed by all the members in your group with how each member contributed to the project. **Please note that this is a group project, while each one of you may have their own contribution, everyone in the group is equally responsible for the final outcome of the project. For instance, if there is a mistake, you cannot simply put the blame on one person who worked on that particular component. Everyone is equally responsible for all content in your final product.**

Project is due by 11:55 pm (ET) on Wednesday April 2nd. One member from the group must submit three documents in Canvas: your group report (as a single PDF), R/Rmd script file, and the e-signed contribution disclosure. The disclosure needs to e-signed by all members of the group by typing their full name. Please do not email me separate R codes from each member in the group. Make sure your R script file is properly consolidated and organized, and include comments so that I can understand without having to contact group members.